Automatic Content Analysis
in
Information Retrieval

G. Salton

Technical Report
No. 68-5

January 1968

Department of Computer Science
Cornell University, Ithaca, N. Y.

Automatic Content Analysis in Information Retrieval

G. Salton

Department of Computer Science
Cornell University, Ithaca, N.Y.

Abstract

The content analysis problem is first introduced, and some of the standard analysis procedures used in information retrieval are reviewed. The principal content analysis methods incorporated into the automatic SMART document retrieval system are then briefly examined and their effectiveness for information retrieval is discussed. Included in the system are word stem matching procedures, synonym recognition, phrase recognition, syntactic analysis, statistical term association techniques, and hierarchical expansion methods.

1. Introduction

Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information. An information retrieval system operates on the one hand in conjunction with a stored collection of items, and on the other with a user population desiring to obtain access to the stored items. The system is thus designed to extract from the files those items which most nearly correspond to existing user needs as reflected in requests submitted by the user population. A library storing books, and serving a population of customers is then, among other things, an example of an information retrieval system.

Prepared for the National Conference on Content Analysis, University of Pennsylvania, November 1967.

This study was supported in part by the National Science Foundation under Grant GN-495.

Conceptually, it is possible to reduce the operations of a typical information retrieval system to two main types: information analysis, and information search and retrieval. The former consists normally in identifying each stored item and each search request by assigning to it one or more content indicators designed to reflect the information content, or the property set, which characterizes the given information item. The latter is mainly a matching operation between content indicators attached to stored items and indicators attached to search requests, followed by the retrieval of those items whose content indicators exhibit a sufficiently high degree of similarity with the query indicators. In a library environment, where the stored items are books or documents, the information analysis normally produces for each item one or more classification numbers, or, alternatively, one or more keywords or index terms, and the retrieval operation is preceded by a comparison of these sets of classification numbers, or keywords, or terms.

In most operational situations, the content analysis of the stored items and search requests is manually conducted by using for this purpose trained cataloguers and indexers, or trained subject experts. The aim of the analysis is to pick, for each item, some set of identifiers which best reflects the interests of the expected user population. Obviously, several different types of indexing strategies can be picked: the content analysis may be quite exhaustive, resulting in a large set of quite specific content indicators, or, alternatively, the analysis may be less detailed, resulting in a smaller set of more general indicators. In the former case, the corresponding retrieval system is likely to produce high precision (but low recall) in that most retrieved items will, in fact, be pertinent to the given query (but some pertinent ones may not be retrieved at all); in the latter case, high recall (but low precision) may result since a search might then produce most everything that is relevant (together with many items that may not be). Compromises are generally made in picking an indexing strategy so as to obtain both a reasonably high recall while holding the precision to within tolerable limits.

While the information analysis is generally performed manually, the information search (that is, the comparison between analyzed items and analyzed search requests) is often mechanized in that a computerized file of stored items is automatically searched and retrieved items are displayed without manual intervention.

The experimental SMART document retrieval system differs from most presently operating systems in that both the analysis and the search and matching operations are performed automatically. Specifically, documents and search requests are stored as abstracts, paragraphs, or sentences in English, and automatic language analysis procedures are used to generate the content identifiers for each stored item. The search and retrieval operations are also conducted automatically by comparing the respective sets of content identifiers for stored documents and incoming search requests.

In the remainder of this study, the content analysis problem arising in information retrieval is briefly outlined. Thereafter, the principal analysis operations incorporated into the SMART system are examined, and their effectiveness in a retrieval environment is described.

An attempt is made to contrast the automatic procedures with alternative methods used in manual systems.

2. The Content Analysis Problem

Before describing the operations of the SMART system, it is useful to introduce a distinction between two different types of automatic text processi systems, the text inference systems and the text retrieval systems. A text inference system is one where one or more written texts are studied, for their own sake so to speak, with the aim of confirming or denying a previously established hypothesis. For example, the hypothesis might be that a given text of unknown authorship was, in fact, written by author A or author B; such a hypothesis might be confirmed by comparing the unknown text with texts known to have been written by A, or by B. Alternatively, the hypothesis might be that the verbal utterances contained in a given written transcript might have been made by a schizophrenic person rather than a normal one; or, that a given political manifesto reflects the Republican Party platform more closely than the Democratic one. In each case, an investigator has made some hypothetical guesses, and a study of the corresponding texts is used to supply evidence from which the truth or falsity of the guesses can be inferred.

In a text retrieval system, on the other hand, the texts are not normally studied for their own sake, but instead they constitute a commodity which is to be distributed to a given user population on demand. No hypothesis is formulated in advance, and the texts are analyzed only in order to determine whether they fit the user's description of what he want:

Superficially, the two types of systems are rather similar, since the same operations are used in both systems as seen in the flowchart of Fig. 1. Search requests, or hypothetical statements, are introduced; then are transformed into components acceptable to the system, often by using authority lists, or dictionaries of various kinds. The resulting sets of content indicators (termed "concept vectors" in Fig. 1) are then compared with the content indicators of a given information store, and items are extracted from the store if the respective content indicators assigned to the stored items match the query indicators sufficiently well.

In actual fact, fundamental distinctions exist between text inference and text retrieval systems as outlined in Table 1. In a text inference system, the users of the system have a well-defined aim, which is generally known in advance; as a result, specialized procedures can be called into play which are specifically chosen to handle the particular problem at hand. In a text retrieval system, on the other hand, the user class is often much larger, and in any case, much more heterogeneous, and its particular interests and concerns are not generally known. A text retrieval system must then be prepared to perform a content analysis which produces acceptable results for a wide class of users of diverse interests, whereas an inference system can attack the problem at hand much more directly.

As a result, the content analysis tools of the inference system, including dictionaries, tables, and thesauruses, can often be tailored to the specific problem in that they reflect the investigators theories relating to the available verbal data [1]. In the retrieval situation, no theories may be said to exist, other than a general knowledge of the language structure, and the dictionaries then take the form of general language processing tools, designed to recognize the regularities which exist over relatively wide ranges of texts in a given subject area. The analysis in the latter case may then be likened to a type of macro-analysis, where an attempt is made to recognize the more important synonymous elements, and the principal subject-verb-object construction of the text, as opposed to a micro-analysis which would use each available word or particle.

The main content analysis operations incorporated into the SMART document retrieval system are outlined in the next few sections, and differences with a typical mechanized text inference system, such as the General Inquirer, are pointed out. [1,2,3,4]

3. The SMART System

SMART is a fully-automatic document retrieval system operating on the IBM 7094. Unlike other computer-based retrieval systems, the SMART system does not rely on manually assigned key words or index terms for the identification of documents and search requests, nor does it use primarily the frequency of occurrence of certain words or phrases included in the texts of documents. Instead, an attempt is made to go beyond simple word-matching procedures by using a variety of intellectual aids in the form of synonym dictionaries, hierarchical arrangements of subject identifiers, statistical and syntactic phrase generation methods and the like, in order to obtain the content identifications useful for the retrieval process.

Stored documents and search requests are then processed without any prior manual analysis by one of several hundred automatic content analysis methods, and those documents which most nearly match a given search request are extracted from the document file in answer to the request. The system may be controlled by the user, in that a search request can be processed first in a standard mode; the user can then analyze the output obtained and, depending on his further requirements, order a reprocessing of the request under new conditions. The new output can again be examined and the process iterated until the right kind and amount of information are retrieved. [5]

SMART is thus designed as an experimental automatic retrieval system of the kind that may become current in operational environments some years hence. The following facilities, incorporated into the SMART system for purposes of document analysis may be of principal interest:

a) a system for separating English words into stems and affixes (the so-called suffix 's' and stem thesaurus methods) which can be used to construct document identifications consisting of the stems of words contained in the documents;

b) a synonym dictionary, or thesaurus, which can be used to recognize synonyms by replacing each word stem by one or more "concept" numbers; these concept numbers then serve as content identifiers instead of the original word stems;

c) a hierarchical arrangement of the concepts included in the thesaurus which makes it possible, given any concept number, to find its "parents" in the hierarchy, its "sons", its "brothers", and any of a set of possible cross references; the hierarchy can be used to obtain more general content identifiers than the ones originally given by going up in the hierarchy, more specific ones by going down, and a set of related ones by picking up brothers and cross-references;

d) statistical procedures to compute similarity coefficients based on co-occurrences of concepts within the sentences of a given collection; the related concepts, determined by statistical association, can then be added to the originally available concepts to identify the various documents;

e) syntactic analysis methods which make it possible to compare the syntactically analyzed sentences of documents and search requests with a pre-coded dictionary of syntactic structures ("criterion trees") in such a way that the same concept number is assigned to a large number of semantically equivalent, but syntactically quite different construction;

f) statistical phrase matching methods which operate like the preceding syntactic phrase procedures, that is, by using a preconstructed dictionary to identify phrases used as content identifiers; however, no syntactic analysis is performed in this case, and phrases are defined as equivalent if the concept numbers of all components match, regardless of the syntactic relationships between components;

g) a dictionary updating system, designed to revise the several dictionaries included in the system:

    i) word stem dictionary
    ii) word suffix dictionary
    iii) common word dictionary (for words to be deleted during analysis)
    iv) thesaurus (synonym dictionary)
    v) concept hierarchy
    vi) statistical phrase dictionary
    vii) syntactic ("criterion") phrase dictionary.

The operations of the system are built around a supervisory system which decodes the input instructions and arranges the processing sequence in accordance with the instructions received. The SMART systems organization makes it possible to evaluate the effectiveness of the various processing methods by comparing the outputs produced by a variety of different runs. This is achieved by processing the same search requests against the same document collections several times, and making judicious changes in the analysis procedures between runs. Illustrations are given of some of the evaluation results obtained with the system when the content analysis methods are covered in more detail in the next few sections. [6]

4. The Stem Dictionaries and Suffix List

One of the earliest ideas in automatic information retrieval was the suggested use of words contained in documents and search requests for purposes of content identification. No elaborate content analysis is then required, and the similarity between different items can be measured simply by the amount of overlap between the respective vocabularies. While such an analysis system is normally considered to be too crude to be of use in a standard text inference system, since no facilities are provided to recognize even the simplest kinds of synonymous constructions, it will be seen that vocabulary matching methods can produce completely satisfactory results for certain types of users of a document retrieval system.

Several different types of entities can be used in a word matching system.

a) the complete English words originally present in documents and search requests can be matched;

b) a minimal amount of vocabulary normalization can be provided by cutting off final 's' endings, so as to confound singular and plural noun forms, and third person verb endings characteristic of standard verb forms;

c) more extensive vocabulary normalization is available if the original text words are first converted to word stem form by deleting standard suffixes and prefixes before matching.

Whichever of the three alternatives is used, the matching process can be applied to all the words in the original text, or alternatively, certain "common" words, deemed to be unimportant as content indicators, can be deleted. Furthermore, the individual text items (words, words with deleted final 's', or word stems) can be weighted in accordance with their presumed importance in a given text. Word frequency is often used as an indication of relative word importance, and a weight proportional to word frequency may then be attached to the text items.

In the SMART system, a decision was made to apply at least a minimal type of language normalization by using word stems instead of original words, deleting common words appearing on an exclusion list, and attaching to each word stem a weight proportional to its frequency in the text.

Stem and suffix dictionaries are first constructed by taking a sample document collection, and using the words occurring in the sample as dictionary entries. New incoming documents and search requests are then processed using a left-to-right letter-by-letter scan in the stem dictionary, and a right-to-left letter-by-letter scan in the suffix dictionary. The longest stem which leaves an acceptable suffix is taken as the correct decomposition of the word. For example, the left-to-right scan of a word like CODING generates potential stems COD (as in CODE), and CODI (as in CODIFY). The latter possibility produces the longer stem, but the remaining suffix NG is not found as an entry in the suffix dictionary. The next longest stem COD is then accepted as correct, since it leaves a proper suffix ING. If a complete word is found in the stem dictionary, the search for the "remaining" suffix is always trivially successful.

A sequence number provided in the stem dictionary is assigned to each acceptable word stem. These sequence numbers are used during subsequent processing to represent the corresponding word stems. The stems pertaining to a given document are then sorted into alphabetic order, and weights are assigned as a function of the corresponding stem frequency. A simplified form of the stem look up is shown in Fig. 2, and a set of sample stem-suffix decompositions is included in Table 2 together with frequency indications and sequence numbers.

A number of English morphological rules are incorporated into the stem-suffix cut-off process to insure that correct stems are identified. Thus, a check is made to identify doubled consonants preceding a suffix (as in HOPPING which is decomposed into HOP+P+ING). Changes from Y to I are also taken into account (as in EASIER which is EASY+ER), as well as deletions of final E before a suffix (as in CODING which is CODE+ING).

The spelling rules actually used during suffix look-up are summarized in Table 3.

A typical search request is shown in Fig. 3 in the form normally used as input to the SMART system. The top part of Fig. 4 shows the reduced form of the text of Fig. 3 as it would appear after look-up in the stem dictionary. It may be seen that a number of common words have been deleted, and weights have been assigned to the remaining entries. (In Fig. 4, a weight of 12 corresponds to an actual frequency of occurrence of 1, and only six characters of each stem are printed, although the complete stem is actually stored).

In a single retrieval system, sets of word stems extracted from documents and search requests can be used directly as an indication of subject similarity. The SMART system does, however, provide more sophisticated procedures to carry out the language analysis. The best-known of these procedures is the standard thesaurus process described in the next section.

5. The Synonym Dictionary or Thesaurus

A thesaurus is a grouping of words, or word stems into certain subject categories, hereafter called concept classes. A typical example is shown in Table 4, where the concept classes are represented by three-digit numbers, and the individual entries are shown under each concept number. In Table 5, a similar thesaurus arrangement is shown in the alphabetic order of the words included. The concept numbers appear in the middle column of Table 5 (concept numbers over 32,000 are attached to "common" words which are not accepted as information identifiers); the last column consists of one or more three-digit syntax codes attached to the words and used for purposes of syntactic analysis.

When constructing a thesaurus for vocabulary normalization, three types of problems must be faced: first, what words should one include in the thesaurus; second, what type of synonym categories should be used (that is, should one aim for broad, inclusive concept classes, or should the classes be narrow and specific); finally, where should each word appear in the thesaurus structure (that is, given a word, what are to be its assigned concept classes).

Obviously, the answers to these questions depend on the use to be made of the thesaurus, and on the environment within which the thesaurus is expected to operate. Experiments conducted with a variety of different types of thesauruses used with the SMART system show that some thesauruses are more effective in a retrieval environment than others. In particular, high-frequency common terms should either be eliminated, or they should appear in concept classes of their own. Low frequency terms should be grouped into classes with other low frequency terms. Terms of little technical significance should be eliminated, and ambiguous terms should appear only in those classes which may be expected to be needed in practice. The thesaurus construction rules used with the SMART system are summarized in Table 6.

A comparison of a typical SMART thesaurus with a thesaurus used with the General Inquirer indicates that the SMART thesaurus classes often have a broader scope, and that many entries normally excluded from a

SMART thesaurus would be used with the General Inquirer. This is true notably of particles expressing negation (which are not used with SMART), of personal pronouns and pronoun references, and of many terms expressing emphasis. The main aim of the two types of thesauruses is, however, the same, namely the transformation of an input text into a set of normalized concepts expressing information content, and both types of thesauruses reflect in one way or another the investigators' theories concerning the language structure and the ways in which words are used to express information content.

In the SMART system, the main entries of the thesaurus are normally word stems, and a text is looked up in the thesaurus one word at a time, in each case replacing the input word by the corresponding thesaurus class or classes. Thus, search requests dealing with the "production of diodes" would normally be assigned the same classes as documents on the "manufacture of transistors". A given text is then transformed into a set of concept numbers with weights, as shown in the middle part of Fig. 4 for the text of Fig. 3.*

The weight of a concept is determined both by the number of words which map into the given concept class, and by the particular thesaurus mapping used. Specifically, a given occurrence of a word is allowed to contribute at most a total weight of 1; the weight of ambiguous words which map into more than one concept is then divided by the number of

---

*It should be noted that the same dictionary look-up program serves for both the stem dictionary and the SMART thesaurus, since the sequence numbers used in the former cannot be distinguished by the computer from the concept classes used in the latter.

applicable concept classes in such a way that a total weight of 1 results (that is, a weight of 1/n is assigned to each concept for a word mapping into n individual concepts). This strategy often results in an automatic resolution of the ambiguities inherent in the vocabulary, because the partial weights of the concepts which actually apply will tend to reinforce each other, whereas the weights of the other inapplicable concepts will be randomly assigned. An example of this phenomenon is shown in Table 7 where the "baseball" category is reinforced with a total weight of 1 and 5/6 for four terms.

A comparison of the first two parts of Fig. 4 shows that the transition from word stem match to thesaurus results in a replacement of stems by concept numbers, and in alterations in the weight structure. For example, a weight of 24 attached to the stem "differen" (from "differential equations") is increased to 36 for the corresponding concept (number 274). The weight is further increased to 72 for phrase concept 379 when phrase assignments are made, as shown in the lower part of Fig. 4.

The philosophy used in the SMART analysis may then be summarized by stating that no attempt is made to eliminate an occasional incorrect concept assignment, but that the automatic procedures are designed to assign a large number of concepts, many of which may be expected to be correctly applicable to the corresponding documents, while at the same time differentiating among individual concepts by the weighting procedure.

The effectiveness of the thesaurus procedure may be judged by the sample evaluation output of Fig. 5 showing recall-precision curves averaged over 17 search requests. <u>Recall</u> is the proportion of relevant material actually retrieved, whereas precision is the proportion of retrieved material actually relevant. For a perfect system which retrieves everything of use to a given customer, and at the same time rejects everything which is not useful, the recall-precision curve would shrink to a single point in the upper right hand corner of Fig. 5 where both recall and precision are equal to 1. In general, the closer the curves are to the upper right hand corner, the better will be the system performance. Fig. 5 which was produced by the evaluation techniques incorporated into the SMART system [6,7] shows that a word stem matching process using only words from the title of documents is clearly inferior to the other methods shown. The word stem match using complete document abstracts is quite effective at the low-recall high-precision end of the curve, where only a few relevant items are desired as output by the user. As more relevant items are wanted, and the recall needs increase, the thesauruses (termed "Harris Two" and "Harris Three" in Fig. 5) becomes increasingly useful. This shows that different types of analysis procedures may be needed to satisfy different types of search requirements. The Harris Three Thesaurus is a recent version of a thesaurus constructed for the field of computer science in accordance with the thesaurus construction principles summarized in Table 6.

6. The Statistical and Syntactic Phrase Dictionaries

Both the thesaurus as well as the stem dictionary are based on entries corresponding either to single words or to single word stems. In attempting to perform a subject analysis of written text, it is possible, however, to go further by trying to locate "phrases" consisting of sets of words which are judged to be important in a given subject area. For example, in the field of computer science, the concepts of "analysis" and "language" may mean many things to many people. On the other hand, the phrase concept which results from a combination of these individual words, that is, "language analysis" has a much more specific connotation. Such phrases can be used for subject identification by building phrase dictionaries to be used in locating combinations of concepts, rather than individual concepts alone. Such phrase dictionaries would then normally include pairs, or triples, or quadruples of words or concepts, corresponding in written texts to the more likely noun and prepositional phrases which may be expected to be indicative of subject content in a given topic area.

Many different strategies can be used in the construction of phrase dictionaries. For example, it is possible to base phrase dictionaries on combinations of high-frequency words or word stems occurring in documents and search requests; alternatively, one may want to use a thesaurus before appeal is made to a phrase dictionary. Furthermore, given the availability of a phrase dictionary one can recognize the presence of phrases in a given text under a variety of circumstances: for example, the existence of a phrase may be recognized whenever the phrase components are present within a given document, or within a sentence of a given document regardless of any actual syntactic relation between the components; alternatively, phrases can be accepted only after verifying that a pre-established syntactic relation actually exists between the phrase components in the document under consideration.

In the SMART system, the phrase dictionaries are based on co-occurrences of thesaurus concepts, rather than text words, in order to profit from the greater degree of language normalization inherent in the use of the concepts. Two principal strategies are used for phrase detection: the so-called statistical phrase dictionary is based on a phrase detection algorithm which takes into account only the statistical co-occurrence characteristics of the phrase components; specifically a statistical phrase is recognized, if all the components are present within a given document, and no attempt is made to detect any particular syntactic relation between the components; on the other hand, the syntactic phrase dictionary includes not only the specification of the particular phrase components which are to be detected, but also information about the permissible syntactic dependency relations which must obtain if the phrase is to be recognized. Thus, if it were desired to recognize the relationship between the concept "program" and the concept "language", then any possible combination of these two concepts such as, for example, "programming language", "languages and programs", "linguistic programs", would be recognized as proper phrases in the statistical phrase dictionary; in the syntactic dictionary, on the other hand, an additional restriction would consist in requiring that the concept corresponding to "program" be syntactically dependent on the concept "language". This eliminates phrases such as "linguistic programs", and "languages and programs", but would permit the phrases "programming languages", or "programmed languages".

A typical excerpt from a statistical phrase dictionary used in connection with the SMART system is shown in Fig. 6. It may be seen that up to six phrase components are permitted in a given phrase, but that the usual phrase specification consists of two, or at most three, components. With each phrase included in Fig. 6, is listed a phrase concept number which replaces the individual component concepts in a given document specification whenever the corresponding phrase is detected by the phrase processing algorithm in use. For example, the first line of Fig. 6 shows that a phrase with concept number 543 is detected whenever the concepts 544 and 608 are jointly present in the document under consideration. Whenever such a phrase concept is attached to a given document specification, the weight of the phrase concept can be increased over and above the original weight of the component concepts to give the phrase specification added importance. This is illustrated in the lower portion of Fig. 4 where the phrase concept 379, representing the concept "differential equations", and obtained by justaposing concepts 274 ("differential") and 181 ("equation"), receives a weight of 72, instead of the original component weights of 36 and 24, respectively.

Since the phrase components used in the SMART system represent concept numbers rather than individual words, a given phrase concept number does then, in fact, represent many different types of English word combinations depending on the number of word stems assigned to each component concept by the original thesaurus mapping.

The syntactic phrase dictionary has a more complicated structure as shown by the excerpt reproduced as Fig. 7. Here, each syntactic phrase, also known as a "criterion tree" or "criterion phrase", consists not only of a specification of the component concepts, but also of syntactic indicators, as well as of syntactic relations which may obtain between the included concepts. For example, the first phrase shown in Fig. 7 carries the concept number 422, and the mnemonic indicator MAGSWI to indicate that this phrase deals in one way or another with magnetic switches. Fig. 7 also shows that the first component of the phrase must consist either of concepts 185 or 624, while the second phrase component must represent concept 225. The indicators after the dollar sign in the output of Fig. 7 carry the syntactic information. In particular, the information given for the phrase MAGSWI indicates that this particular phrase must be either of syntactic types 7, or 15, or 16.

The automatic process used to perform the syntactic analysis of the original tests and to assign syntactic indicators to the concepts, as well as the matching process between syntactic phrases occurring in documents and search requests have previously been described in the literature. [8] An evaluation of the phrase techniques shows that the statistical phrase process is often more effective than a simple thesaurus look-up. On the other hand, the automatic syntactic procedures appear to be not substantially superior to the statistical methods, even though they are far more expensive to perform on the computer. The reasons for this unexpected result may be due in part to the relative inadequacy of presently existing programs for automatic syntactic analysis, and in part to the face that the syntactic procedures appear to be too refined for the document retrieval environment in which they are used. Thus, in the sentence "for people who need a great deal of information, effective retrieval is vital", the phrase "information retrieval" would not be recognized by the syntactic procedures in use, since "information" and "retrieval" do not exhibit the appropriate syntactic relationships. The sentence does, however, deal with information retrieval, a fact which is properly recognized by the statistical phrase methods used. This example demonstrates again that a content analysis method which is too sophisticated is not more useful than one which is not sophisticated enough. The difficulty lies in recognizing the appropriate depth of the analysis to be used in each given case.

7. The Concept Hierarchy

Hierarchical arrangements of subject headings have been used for many years in library science and related documentation activities. In general, such arrangements make it possible to classify more specific topics under more general ones, and to formulate a search request by starting with a general formulation, and progressively narrowing the specification down to those areas which appear to be of principal interest.

In a content analysis system, a hierarchical arrangement of words or word stems can be used both for information identification and for retrieval purposes. Thus, if a given search request is formulated in terms of "syntactic dependency trees", and it is found that not enough useful material is actually obtained, it is possible to "expand" this request to include all "tree structures" or indeed all "abstract graphs", by using a hierarchical subject classification.

A hierarchy of concept numbers rather than text words is included in the SMART system, and it is assumed that a thesaurus look-up operation precedes any hierarchical expansion operation. A typical example from the SMART concept hierarchy is shown in Fig. 8. The broad, more general concepts appear on the left side of the figure, corresponding to the "roots" of the hierarchical tree; and the more specific concepts appear further to the right. For example, concept 270 is the root of a sub-tree, this concept has four sons on the next lower level, namely concepts 224, 471, 472, and 488. Concept 224 is turn has two sons, labelled 261 and 331; similarly, concept 471 has four sons, including 338, 371, 458, and 470. It may be seen from Fig. 8, that the sons of a concept, representing more specific terms, are shown below their parents and further to the right.

The hierarchy of Fig. 8 also provides for the inclusion of cross references from one concept to another, connected to the original concept by broken lines. Such cross references represent general, unspecified types of relations between the corresponding concepts, and receive in general a different interpretation than the generic inclusion relations normally represented by the hierarchy.

It would be nice if it were possible to give some generally applicable algorithm for constructing hierarchical subject arrangements. This is, in fact, a topic which has preoccupied many people including mathematicians, philosophers, and librarians for many years. In general, one might expect that broad concepts should be near the top of tree (close to the root), whereas specific concepts should be near the bottom (close to the leaves); furthermore, there appears to be some relationship between the frequency of occurrence of a given concept in a document collection, and its place in the hierarchy. More specifically, those concepts which exhibit the highest frequency of occurrence in a given document collection, and which by this very fact appear to be reasonably common, should be placed on a higher level than other concepts whose frequency of occurrence is lower.

Concerning the specific place of a given concept with the hierarchy, this should be made to depend on the user population and on the type of expansion which is most often requested. Thus, a concept corresponding to "syntactic dependency tree' would most reasonably appear under the broader category of 'syntax", which in turn could appear under the general class of "language", assuming that the user population consists of linguists or grammarians; on the other hand, if the users were to be mathematicians or algebraists, then the "syntactic dependency trees" should probably appear under "abstract trees", which in turn would come under "graph theory", a branch of algebra. It does not appear reasonable to expect that a hierarchical arrangement of concepts will serve equally well for all uses under all circumstances. Rather any hierarchy will serve its function, if it can be counted upon to suggest ways of broadening or narrowing a given search request or a given interpretation of the subject matter under most of the circumstances likely to arise in practice.

8. Statistical Term Associations

The content analysis procedures described up to now either do not take into account any kind of relationships between individual content identifiers, or, alternatively, the relationships that exist are specified by the dictionaries used in the analysis. The phrase dictionaries, for example, specify a type of association between individual concepts within a phrase, and the hierarchical expansion operations make use of generic inclusion relations between concepts.

It is, however, also possible in a retrieval environment to take into account various kinds of associations between concepts which are inherent in the query and document texts, instead of being specified by a dictionary or thesaurus. Specifically, if it is assumed that two document identifiers are related whenever they are found to co-occur frequently in the same context — for example, in the same sentences of a document, or in the same documents of a collection — then it is possible to compute an index of similarity between each pair of concepts based on these co-occurrence characteristics. Thereafter, each given concept vector representing a document or search request, can be expanded by addition of all the associated concepts whose similarity coefficients with some original concept are sufficiently high.

Consider as an example, a typical set of concept vectors such as those shown in Fig. 9, for eight documents. The terms assigned to the eight documents are labelled A to F, and no weights are used in the example of Fig. 9. A similarity coefficient can now be computed between each pair of terms, based on joint assignment of the corresponding terms to the documents of the collection, by comparing the corresponding two columns of the matrix of Fig. 9. If the similarity coefficient between two terms is computed by a formula such as

$$\frac{\text{Number of joint occurrences of terms i and j}}{\text{Number of i's + Number of j's - Number of joint occurrences}}$$

then the similarity matrix of Fig. 10 results for the original specification of Fig. 9. To compute, for example, the similarity between terms A and B, a comparison of the first two columns of Fig. 9 shows that there exist two joint assignments (to documents 2 and 8), four individual occurrences of A (documents 1, 2, 5, and 8), and four occurrences of B (documents 2, 4, 7, and 8). The similarity coefficient is then

$$\text{Similarity (A,B)} = \frac{2}{4+4-2} = 2/6$$

If the further assumption is now made that a similarity coefficient of at least 2/5 is to be indicative of a statistical association between the corresponding terms, then the four pairs (A,D), (B,F), (C,E), and (C,F) would be accepted as associated, according to the statistical criterion used. Consequently, to a document specification consisting of terms A, B, and C, one might then add the associated terms D, E, and F, thus ensuring, hopefully, that a query dealing with "airplanes" would also retrieve documents about "aircraft".

Tests were made to determine to what extent the automatically generated term association methods incorporated into the SMART system could be considered to be equivalent to the manually or semi-automatically constructed thesauruses [6,9]. The results indicate that while a retrieval system using term associations provides greater effectiveness than one based on simple word matching alone, the normal thesaurus process is much more effective as a language normalization device than the statistical word associations. Furthermore, the associations which are automatically determined are not related to those specified in the thesaurus, and do not approximate normal synonym relations between words. Associative methods are therefore most effective in situations where a thesaurus is not available, and where the time and effort needed to generate one cannot be expended.

9. User-Controlled Information Search

In the SMART system, document retrieval takes place following the information analysis. Specifically, the concept vectors which are generated for the individual documents during the analysis phase are compared with the concept vectors assigned to the search requests, and those documents which are found to be most similar to the queries are retrieved for the user's attention. A typical output form is shown in Fig. 11 in the format in which it is transmitted to the user. The original query (already shown in Fig. 3) is reproduced at the top of the figure, followed by an itemization of the first few documents in decreasing correlation order with the search request. The user now has the option to quit, or to request that additional items be displayed for his attention.

While the SMART system includes a large variety of content analysis procedures, which produce different types of results for different users — some stressing high recall and some high precision — it is unreasonable to expect wholly satisfactory service to all users under all circumstances, particularly if only a single search is made of the stored collection. Attempts to meet the user problem usually take the form of multiple rather than single searches. Thus, instead of submitting a search request and obtaining in return a final set of relevant items, a partial search is made first and, based on the preliminary output obtained, the search parameters are adjusted before attempting a second, more refined search. The adjustments made may then be different from user to user, depending on individual needs, and the search process may be repeated as often as desired. A typical user feedback system is shown in simplified form in Fig. 12.

Several strategies are available for improving the results of a search, as summarized in Table 8. [4] The first is simply a <u>mechanized dictionary</u> print-out routine in which a set of potential search terms, related to those initially used by the requestor, are extracted from the stored dictionary and presented to the user. The user is then asked to reformulate the original query after selecting those new associated terms which appear to him to be most helpful in improving the search results. Typically, the statistical term associations previously discussed can be used to obtain the set of related terms, or the sets of associated thesaurus classes can be taken from the thesaurus. This search optimization procedure is straight-forward, but leaves the burden of rephrasing the query in the user's hands.

A second strategy consists in automatically modifying a search request by using the partial results from a previous search. Specifically, the user is asked to examine the documents retrieved by an initial search, and to designate some of them as either relevant (R) or irrelevant (N) to his purpose. Concepts from the documents termed relevant can then be added to the original search request if not present already, or their importance can be increased by a suitable adjustment of weights; contrariwise, terms from documents designated as irrelevant can be deleted or demoted. A great deal of work has been done to optimize this kind of relevance feedback operation, and evaluation results appear to indicate that the process produces considerable improvements in search effectiveness. [5]

The third possibility for search optimization leaves the search request effectively unchanged but alters the analysis process. This requires a retrieval organization, like SMART, providing several possible content analysis techniques and an iterative search procedure which can utilize the various analysis methods for retrieval purposes. User feedback can also serve here as a basis for choosing from among the large number of available analysis methods the one which seems most appropriate in each given case.

10. Summary

The SMART system is a fully-automatic text processing system which includes a large variety of content analysis methods designed to transform incoming English texts into sets of concept vectors reflecting information content. The programs consist of about 150,000 program steps on an IBM 7094/II, and fourteen tapes are required on-line if the full facilities of the system are called into play. Various parts of the system have been reprogrammed for an IBM 360/65, but the full system is not yet available for the 360 at the time of this writing.

The SMART system is implemented as a text retrieval system; however, given appropriate inputs and dictionaries, it could operate just as easily as a text inference system. In fact, all the facilities provided by the programs which constitute the General Inquirer are also present in SMART, including dictionary look-up programs, "tag tally" programs, and syntactic analysis facilities. The SMART programs in addition permit a fully-automatic text analysis without manual operations at the input side, and include also sophisticated search and retrieval evaluation facilities. It is to be hoped that sooner or later the SMART programs may find application in the social sciences for content analysis and other related research endeavors.

References

[1] P. J. Stone, D. C. Dunphy, M. S. Smith, D. M. Ogilvie, The General Inquirer — A Computer Approach to Content Analysis — Studies in Psychology, Sociology, Anthropology, and Political Science, M.I.T. Press, 1966.

[2] P. J. Stone, A Computer Approach to Content Analysis: Studies Using the General Inquirer System, Proceedings of the Spring Joint Computer Conference, Spartan Books, Vol. 23, April 1963.

[3] G. Salton and M. E. Lesk, The SMART Automatic Document Retrieval System — An Illustration, Communications of the ACM, Vol. 8, No. 6, June 1965.

[4] G. Salton, Progress in Automatic Information Retrieval, IEEE Spectrum, Vol. 2, No. 8, August 1965.

[5] G. Salton, Search Strategy and Optimization of Retrieval Effectiveness, FID-IFIP Conference on Mechanized Storage, Retrieval and Dissemination, Rome, June 1967.

[6] G. Salton and M. E. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 14, No. 1, January 1968.

[7] G. Salton, Evaluation of Computer-Based Retrieval Systems, Proceedings of the FID Congress 1965, Spartan Books, 1966.

[8] G. Salton, Automatic Phrase Matching, in Readings in Automatic Language Processing, D. Hays, editor, American Elsevier Publishing Co., 1966.

[9] M. E. Lesk, Word-word Associations in Document Retrieval Systems, Report No. ISR-13 to the National Science Foundation, Section IX, Department of Computer Science, Cornell University, December 1967.

	Text Inference	Text Retrieval
Search Request	Hypothesis made by one or more investigators	Requests for information made by a given user population
User Population	Investigators interested in test study	Customer population desiring access to information store
Query Type	Generally oriented in a specific direction	Often unknown in advance
Content Analysis	Specialized toward solution of a given problem	General in order to fit heterogenous user population

Principal Differences between Text Inference
and Text Retrieval Systems

Table 1

Frequency	Stem	Suffix	Sequence Number	Frequency	Stem	Suffix	Sequence Number
11	MODULE	S	2099	12	CONCEPT		2113
11	PLACE	S	2100	12	DECIS	ION	2114
11	RESPONCE		2101	12	DEPOSIT	ED	2115
11	RF		2102	12	DUE		2116
11	SOURCE		2103	12	ECONOM	ICAL	2117
11	THICK		2104	12	ESAKI		2118
11	TRUNC	ATION	2105	12	EXAMIN	ED	2119
11	WAVE		2106	12	FUNCTION	AL	2120
11	WHEREB	Y	2107	12	GRAPH		2121
11	WIR	ING	2108	12	HAV	ING	2122
12	ALPHABET	ICAL	2109	12	IMPROVE	MENT	2123
12	BASE		2110	12	IMPROV	ED	2124
12	CAP	ABLE	2111	12	INDIVIDU	AL	2125
12	CENT		2112	12	LEAST		2126

Excerpt from Word Stem Frequency List

Table 2

Input Word	Corresponding Thesaurus Stem	Original Stem Detected	Modified Stem	+	Suffix
HOPPING	HOP	HOPP	HOPP	+	ING
HOPING	HOPE	HOP	HOP	+	ING
HANDING	HAND	HAND	HAND	+	ING
EASIER	EASY	EAS	EASI	+	ER
EASING	EASE	EAS	EAS	+	ING

Spelling Rules Incorporated in the
Stem Look-up Process

Table 3

408 DISLOCATION JUNCTION MINORITY-CARRIER N-P-N P-N-P POINT-CONTACT RECOMBINE TRANSITION UNIJUNCTION	411 COERCIVE DEMAGNETIZE FLUX-LEAKAGE HYSTERESIS INDUCT INSENSITIVE MAGNETORESISTANCE SQUARE-LOOP THRESHOLD
409 BLAST-COOLED HEAT-FLOW HEAT-TRANSFER	412 LONGITUDINAL TRANSVERSE

Thesaurus Excerpt (Concept Class Order)

Table 4

Text Words	Concept Numbers	Syntax Codes
BLOCK	663	070043040
BLUEPRINT	58	070043
BOMARC	324	070
BOMBARD	424 0343	043
BOMBER	346	070
BOND	105	070043
BOOKKEEPING	34	070
BOOLEAN	20	001
BORROW	28	043
BOTH	32178	008080012
BOUND	523 0105	070043134135
BOUNDARY	524	070
BRAIN	404 0235	070
BRANCH	48 0042	070042
BRANCHPOINT	23	070
BREAK	380	043040070
BREAKDOWN	689	070
BREAKPOINT	23	070
BRIDGE	105 0458 0048	070043
BRIEF	32232	001043071
BRITISH	437	001071
BROAD-BAND	312	001071

Thesaurus Excerpt in Alphabetic Order

Table 5

Type of Term	Thesaurus Rule
Very rare terms	do not place into separate categories in the thesaurus, but combine if possible with other rare terms to form larger classes (low frequency categories provide few matches between stored items and search requests)
Very common terms	high-frequency terms should be either eliminated since they provide little discrimination, or should be placed into synonym classes of their own so as not to submerge other terms with which they might be grouped
Terms of no technical significance	terms which have no special significance in a given technical area (such as "begin", "automatic", "system", etc. in the computer science area) should be excluded from the thesaurus
Ambiguous terms	ambiguous terms should be entered into the thesaurus only in those senses likely to occur in the given subject area

Sample Thesaurus Construction Rules

Table 6

		Thesaurus Classes					Thesaurus Frequency
		Lamps	Games Baseball	Animals	Military Usage	Clothing	
Original Terms	base	1/3	1/3		1/3		1
	bat		1/2	1/2			1
	glove		1/2			1/2	1
	hit		1/2		1/2		1
Frequency in Document		1/3	1-5/6	1/2	5/6	1/2	

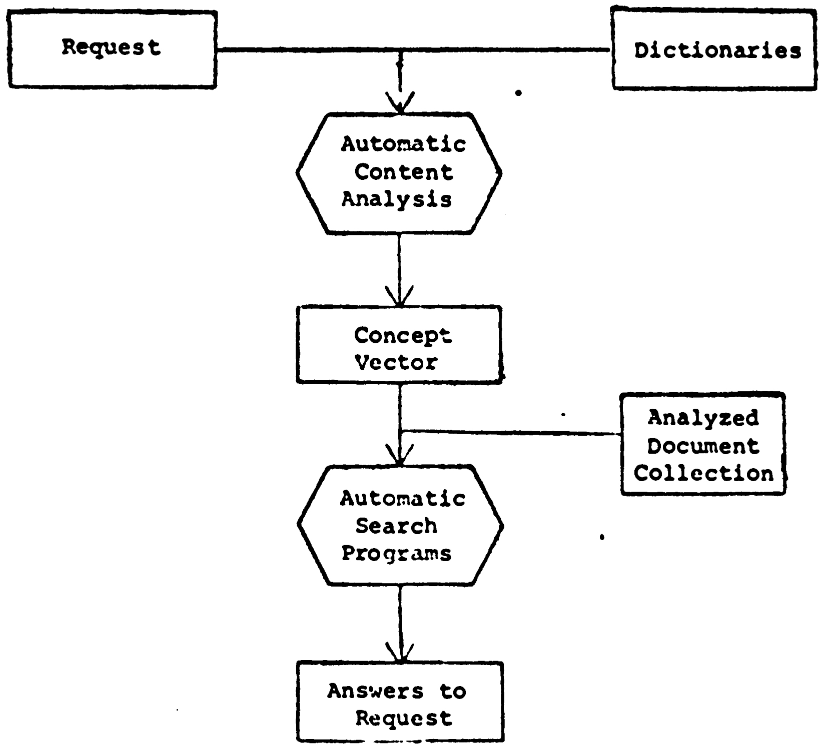
Sample Thesaurus Mapping

Table 7

Type of System Alteration	User Function
Display of related terms from stored dictionary or generation of term associations	User adds to search request related terms suggested to him by the system
Automatic query change by promoting terms from relevant documents and demoting these from irrelevant documents	User identifies some previously retrieved items as either relevant or irrelevant to his purposes
Automatic change of search process, using additional dictionaries, hierarchies, or statistical and syntactic analysis methods	User criticizes result of an initial search by pointing out insufficiency of output (narrow or broad subject interpretation, theme recognition, and so on)

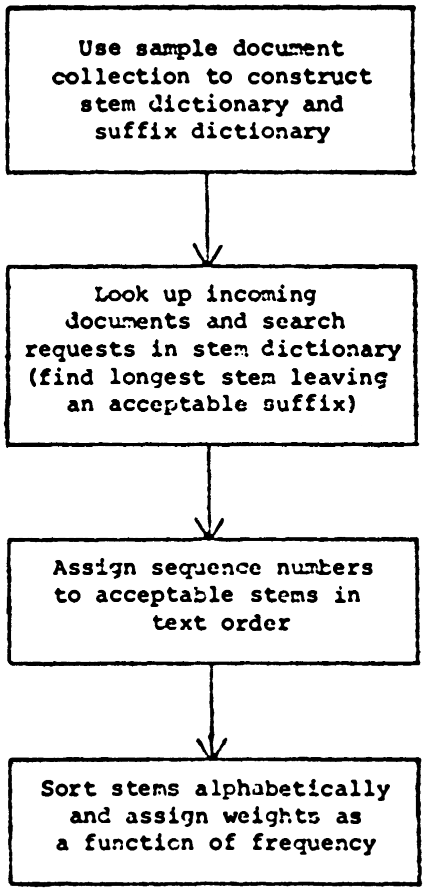
Methods for Search Optimization

Table 8



Simplified Text Processing System

Fig. 1



Simplified Look-up in Stem Dictionary

Fig. 2

ENGLISH TEXT PROVIDED FOR DOCUMENT DIFFERNTL EQ SEPT. 28, 1964

GIVE ALGORITHMS USEFUL FOR THE NUMERICAL SOLUTION	1
OF ORDINARY DIFFERENTIAL EQUATIONS AND PARTIAL DIFFER-	1
ENTIAL EQUATIONS ON DIGITAL COMPUTERS. EVALUATE THE	1
VARIOUS INTEGRATION PROCEDURES (TRY RUNGE-KUTTA,	2
MILNE-S METHOD) WITH RESPECT TO ACCURACY, STABILITY,	2
AND SPEED.	2

TYPICAL SEARCH REQUEST

Fig. 3

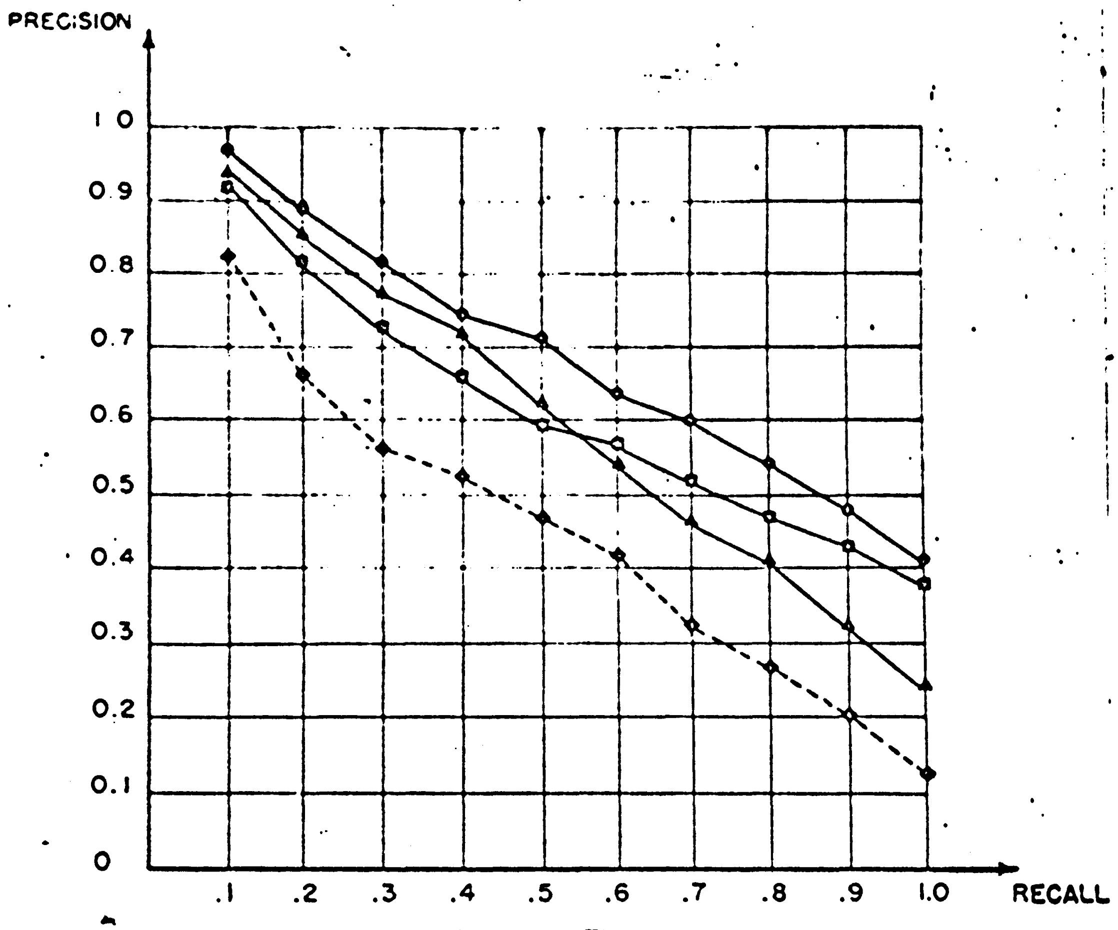
OCCURRENCES OF CONCEPTS AND PHRASES IN DOCUMENTS

DOCUMENT	CONCEPT, OCCURS					
DIFFERNTL EQ	ACCUR 12	ALGORI 12	COMPUT 12	DIFFER 24	DIGIT 12	
	EQU 24	EVALU 12	GIVE 12	INTEGR 12	METHOD 12	STEM
	NUMER 12	ORDIN 12	PARTI 12	PROCED 12	RUNGE- 12	DICTIONARY
	SOLUT 12	SPEED 12	STABIL 12	USE 12	VARIE 12	
DIFFERNTL EQ	4EXACT 12	8ALGOR 12	13CALC 18	71EVAL 6	92DIGI 12	REGULAR
	110AUT 12	143UTI 12	176SOL 12	179STD 12	181QUA 24	THESAURUS
	269ELI 4	274DIF 36	356VEL 12	357YAW 4	384TEG 12	
	428STB 4	505APP 24				
DIFFERNTL EQ	4EXACT 12	8ALGOR 12	13CALC 18	71EVAL 6	92DIGT 12	STATISTICAL
	110AUT 12	143UTI 12	176SOL 12	179STD 12	181QUA 24	PHRASES
	269ELI 4	274DIF 36	356VEL 12	357YAW 4	375NUM 36	LOOK-UP
	379DIF 72	384TEG 12	428STB 4	505APP 24		

INDEXING PRODUCTS FOR "DIFFERENTIAL EQUATIONS"

Fig. 4

Stem Title Only		Word Stem Match		Thesaurus Harris Two		Thesaurus Harris Three	
0.1	0.8307	0.1	0.9563	0.1	0.9551	0.1	0.9735
0.2	0.6800	0.2	0.8648	0.2	0.8242	0.2	0.8973
0.3	0.5720	0.3	0.7986	0.3	0.7389	0.3	0.8245
0.4	0.5323	0.4	0.7381	0.4	0.6796	0.4	0.7551
0.5	0.4816	0.5	0.6371	0.5	0.6070	0.5	0.7146
0.6	0.4142	0.6	0.5589	0.6	0.5702	0.6	0.6499
0.7	0.3489	0.7	0.4877	0.7	0.5233	0.7	0.6012
0.8	0.2687	0.8	0.4086	0.8	0.4821	0.8	0.5514
0.9	0.2016	0.9	0.3426	0.9	0.4452	0.9	0.4973
1.0	0.1463	1.0	0.2613	1.0	0.3951	1.0	0.4118



Comparison Based on Thesauruses
(averages over 17 search requests)

Fig. 5

Phrase Concept	Component Concepts				
543	544	608	-0	-0	-0
282	280	281	-0	-0	-0
282	306	281	-0	-0	-0
280	69	648	-0	-0	-0
280	69	215	-0	-0	-0
694	1285	1284	-0	-0	-0
291	265	290	-0	-0	-0
291	265	496	-0	-0	-0
422	646	185	-0	-0	-0
640	309	290	-0	-0	-0
294	21	293	-0	-0	-0
393	21	635	-0	-0	-0
393	635	106	-0	-0	-0
294	21	245	-0	-0	-0
695	44	150	-0	-0	-0
78	572	565	-0	-0	-0
411	370	328	-0	-0	-0
411	370	389	-0	-0	-0
411	370	476	-0	-0	-0
666	46	601	-0	-0	-0
666	330	53	601	-0	-0
666	347	46	-0	-0	-0

Excerpt from Statistical Phrase Concept Dictionary

Fig. 6

```
Name Output  First Second Type 7 Type 15 Type 16
 of   Con-    Node  Node  Serial Serial  Serial
Tree  cept     Concepts    143    143     399

MAGSWI=422(185,624)/(225)$7/143,15/398,16+
MANMCH=517(600)/(516)$7/144,15/400
MANROL=286(290)/(113)$7/145,5+,15/401,16+,19+
MATHOP=594(615)/(7,116,376)$7/147
MCHBKD=69(689)/(600)$1/148
MCHCOD=304(102,281)/(14,41,600,601)$1/149,15/404
MCHOPD=93(615)/(600)$7/150
MCHORI=41(513)/(600,601)$7/151,15/405
MCHTIM=691(617)/(52,600,601,605,1231)$7/152
MCHTIM=691(617)/(72,615)$1/153
MCHTRA=303(98)/(119,600)$1/154,4+,5+,6+,10+,15/406,16+,19+
MEMACC=593(672)/(121)$1/159,15/409
MEMCOR=557(669)/(121)$7/137,15/395
MEMEFF=294(64)/(121)$1/160,6+,15/410
MEMSPA=552(212)/(121)$1/162,13+,15/411
```

Excerpt from Syntactic Criterion Phrase Dictionary

Fig. 7

							Concept Number	Sequence Number
053		350		625			53	1
	584						584	2
							-0	3
130							130	4
	074		192				74	5
	114		725		101		114	6
	494		725		101		494	7
							-0	8
195							195	9
	246		120				246	10
	374		120				374	11
	468						468	12
	469						469	13
	491						491	14
							-0	15
260							260	16
	485		435				485	17
							-0	18
270							270	19
	224						224	20
		261		130			261	21
		331					331	22
	471		641				471	23
		338					338	24
		371					371	25
		458					458	26
		470					470	27
	472		641		200		472	28
		034		641			34	29
	488						488	30
							-0	31
309		597		321			309	32
	551		341		335		551	33
	628		659		597	630	628	34
	642		341				642	35
	643		659				643	36

SONS

CROSS-REFERENCES

Hierarchy Excerpt

Fig. 8

	Terms					
Document 1	A	-	-	D	-	-
Document 2	A	B	-	D	-	-
Document 3	-	-	C	D	E	F
Document 4	-	B	C	-	-	F
Document 5	A	-	-	D	E	-
Document 6	-	-	C	-	E	-
Document 7	-	B	-	D	-	F
Document 8	A	B	C	-	-	-

Original Term Assignment for Eight Documents

Fig. 9

	A	B	C	D	E	F
A	.	2/6	1/7	3/6	1/6	0
B		.	2/6	2/7	0	2/5
C			.	1/8	2/5	2/5
D				.	2/6	2/6
E					.	1/5
F						.

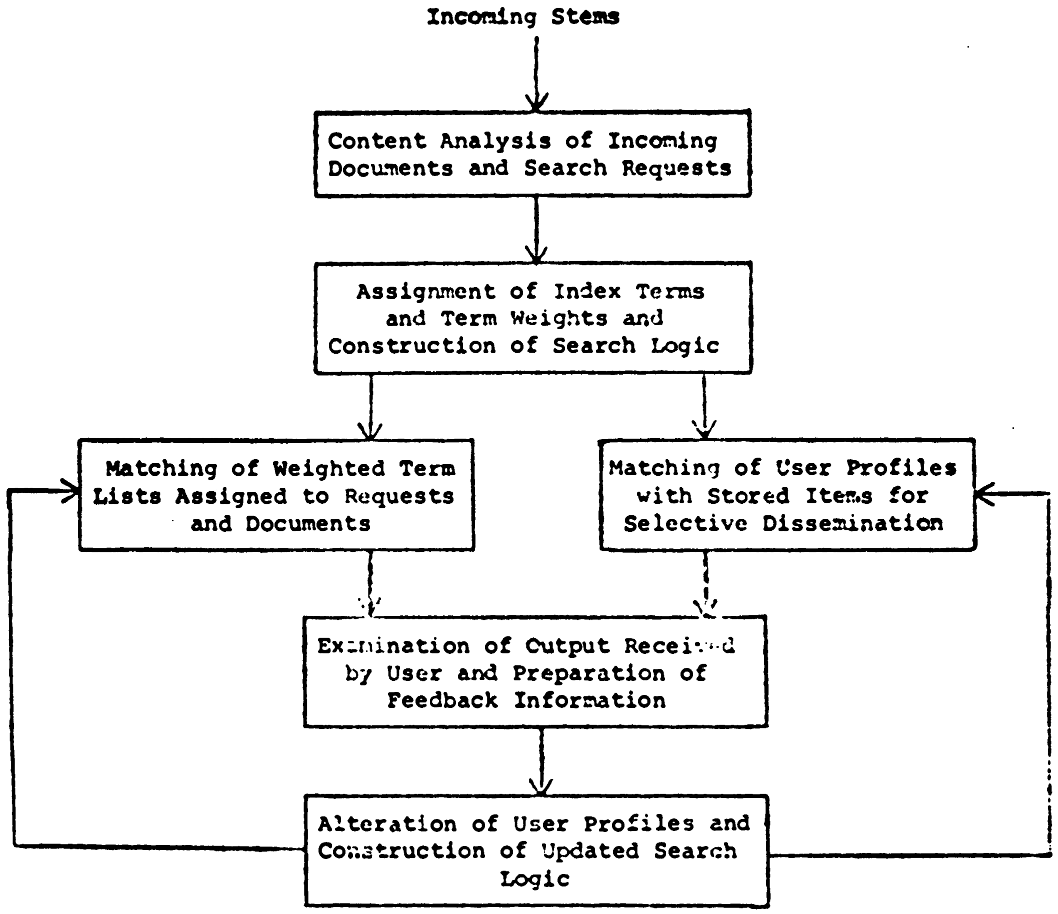
Term-Term Similarity Matrix

Fig. 10

```
ANSWERS TO REQUESTS FOR DOCUMENTS ON SPECIFIED TOPICS          SEPTEMBER 28, 1964    PAGE 83

              CURRENT REQUEST - *LIST DIFFERNL EQ NUMERICAL DIGITAL SOLN OF DIFFERENTIAL EQUATIONS

REQUEST    *LIST DIFFERNTL EQ NUMERICAL DIGITAL SOLN OF DIFFERENTIAL EQUATIONS
-------
           GIVE ALGORITHMS USEFUL FOR THE NUMERICAL SOLUTION OF ORDINARY
           DIFFERENTIAL EQUATIONS AND PARTIAL DIFFERENTIAL EQUATIONS ON DIGITAL
           COMPUTERS . EVALUATE THE VARIOUS INTEGRATION PROCEDURES (E.G. RUNGE--
           KUTTA, MILNE-S METHOD) WITH RESPECT TO ACCURACY, STABILITY, AND SPEED .


ANSWER          CORRELATION          IDENTIFICATION
------          -----------          --------------
384STABILITY     0.6675         STABILITY OF NUMERICAL SOLUTION OF DIFFERENTIAL EQUATIONS
                                W. E. MILNE AND R. R. REYNOLDS (OREGON STATE COLLEGE)
                                J. ASSOC. FOR COMPUTING MACH. VOL 6 PP 196-203 (APRIL, 1959)

ANSWER          CORRELATION          IDENTIFICATION
------          -----------          --------------
360SIMULATIN     0.5758         SIMULATING SECOND-ORDER EQUATIONS
                                D. G. CHADWICK (UTAH STATE UNIV.)
                                ELECTRONICS VOL 32 P 64 (MARCH 6, 1959)

ANSWER          CORRELATION          IDENTIFICATION
------          -----------          --------------
200SOLUTION      0.5663         SOLUTION OF ALGEBRAIC AND TRANSCENDENTAL EQUATIONS ON AN AUTOMATIC
                                DIGITAL COMPUTER
                                G.N. LANCE (UNIV. OF SOUTHAMPTON)
                                J. ASSOC. FOR COMPUTING MACH., VOL 6, PP 97-101, JAN., 1959
ANSWER          CORRELATION          IDENTIFICATION
------          -----------          --------------
392ON COMPUT     0.5508         ON COMPUTING RADIATION INTEGRALS
                                R. C. HANSEN (HUGHES AIRCRAFT CO.), L. L. BAILIN (UNIV. OF SOUTHERN
                                CALIFORNIA, AND R. W. RUTISHAUSER (LITTON INDUSTRIES, INC.)
                                COMMUN. ASSOC. FOR COMPUTING MACH. VOL 2 PP 28-31 (FEBRUARY, 1959)
ANSWER          CORRELATION          IDENTIFICATION
------          -----------          --------------
386ELIMINATI     0.5483         ELIMINATION OF SPECIAL FUNCTIONS FROM DIFFERENTIAL EQUATIONS
                                J. E. POWERS (UNIV. OF OKLAHOMA)
                                COMMUN. ASSOC. FOR COMPTING MACH. VOL 2 PP 3-4 (MARCH, 1959)
```

Answers to Search Requests

Fig. 11



Simplified User Feedback Process

Fig. 12